

## Supplementary Patient and Methods

This Supplementary material is meant to provide more details on the “Patients and Methods” section of the manuscript.

### (i) Structured consensus exercise to develop domains and core set items

This details have been published in the Khanna D, et al. Annals of the Rheumatic diseases 2008; 67(5):703-9 . The original domains and core set items included were

<b>SKIN</b>	
1	Modified Rodnan Skin Score (Range 0-51)
2	Visual analog scale (VAS) or /Likert of <b>patient</b> global assessment for skin activity
3	VAS or /Likert of <b>physician</b> global assessment for skin activity
4	Durometer
<b>MUSCULOSKELETAL</b>	
1	Tender Joint Count
2	Tendon friction rubs assessed by the <b>physician</b>
<b>CARDIAC</b>	
1	<sup>1</sup> Cardiac Echocardiogram with doppler <sup>1</sup>
2	<sup>2</sup> Right heart catheterization*
3	6-minute walk test*
4	Borg Dyspnea Instrument *
<b>PULMONARY</b>	
1	<sup>3</sup> Pulmonary Function Testing
2	<sup>4</sup> Validated measure of Dyspnea
3	Breathing VAS from the Scleroderma Health Assessment Questionnaire (S-HAQ) <sup>42</sup>
4	High resolution computer tomography (HRCT): quantifiable scale*
<b>RENAL</b>	
1	Calculated creatinine clearance based on serum creatinine(Cockcroft-Gault or MDRD formula)
2	Pre-defined renal crisis (Presence or absence)
<b>GASTROINTESTINAL</b>	
1	Body mass index (BMI)
4	Validated gastrointestinal tract VAS scale (part of S-HAQ) <sup>42</sup> or other SSc-Validated GI questionnaire
<b>HEALTH-RELATED QUALITY OF LIFE AND FUNCTION</b>	
1	Health Assessment Questionnaire-Disability Index (HAQ-DI) <sup>43</sup>
2	VAS-pain scale from the HAQ-DI <sup>43</sup>
3	SF-36 version 2 <sup>44</sup>
<b>GLOBAL HEALTH</b>	
1	VAS/Likert <b>patient</b> global severity
2	VAS/Likert <b>physician</b> global severity

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

3	<sup>5</sup> Scleroderma-related health transition by patient
4	<sup>6</sup> Scleroderma-related health transition by physician
<b>RAYNAUD'S</b>	
1	Raynaud's Condition Score <sup>45</sup>
2	VAS Raynaud's (part of S-HAQ) <sup>42</sup>
<b>DIGITAL ULCERS</b>	
1	Active digital tip ulcer count on the volar surface
2	VAS digital ulcer (part of S-HAQ) <sup>42</sup>
<b>BIOMARKERS</b>	
1	Acute phase reactant(s)—Erythrocyte sedimentation rate and/or C-reactive protein

† Standardized central reading mechanism strongly encouraged, \* if relevant to the study

Parameters to be measured for the items in the core set

- <sup>1</sup>Echocardiogram with Doppler— Measure pulmonary artery systolic pressure, left and right ventricular systolic and diastolic parameters, pericardial effusion, and chamber size
- <sup>2</sup>Right heart catheterization— pulmonary artery pressure, pulmonary vascular resistance, wedge pressure, cardiac output/cardiac index
- <sup>3</sup>Pulmonary Function Test— Forced Vital Capacity (FVC), Diffusion Capacity (DLCO), Total Lung Capacity (TLC)
- <sup>4</sup>Validated dyspnea questionnaires e.g., University of California San Diego (UCSD) Dyspnea Questionnaire, St. George Respiratory Questionnaire, Mahler's Dyspnea Index
- <sup>5</sup>Scleroderma-related health transition by patient—Compared to one year ago, how do you rate your overall scleroderma— much better, a little better, no change, a little worse, much worse)
- <sup>6</sup>Scleroderma-related health transition by physician (Compared to one year ago, how do you rate your patient's overall scleroderma— much better, a little better, no change, a little worse, much worse)

**(ii) Data collection and evaluation of psychometric properties in a longitudinal observational cohort.**

Thirty-one core set items, based on a structured consensus exercise (identified in (i)) were prospectively collected in 200 patients with early diffuse cutaneous systemic sclerosis (dcSSc) at 4 U.S. Scleroderma centers. Core set items included continuous, binary, and categorical variables.

We assessed the feasibility, reliability, face and content validity, redundancy and sensitivity to change of each core set item using the following criteria:

- *Feasibility*: a core set item was defined to be feasible is more than 50% of the subjects had complete data at both baseline and 12-month follow-up.
- *Reliability*: Internal consistency reliability for the SF-36, the St. George's questionnaire and the HAQ-DI questionnaires was evaluated using Cronbach's alpha and found to be acceptable  $\geq 0.70$ .
- *Face and content validity*: This is based on the consensus exercise in step (i).

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

- *Redundancy:* To determine redundancy in the core set items we computed the correlation coefficients among the 31 core set items at baseline, at the 12-months follow up and in the change from baseline to follow up. Correlations were computed using Spearman or Pearson correlation coefficient depending on the core set item: Spearman correlation was used when one or both core set items were discrete (e.g. either binary or categorical); Pearson correlation was used when core set items were continuous. A correlation coefficient, either Spearman or Pearson, of  $\geq 0.80$  at either baseline or 12-month follow-up indicated core set items that were redundant.
- *Sensitivity to change:* We determined sensitivity to change for each core set item using two transition questions using a patient and a physician question. At the 1-year follow-up, each subject and clinician was asked to rate the overall scleroderma condition compared to the previous year. Both transition questions used a Likert 1-5 scale with 1 representing “much better” and 5 indicating “much worse”. Those who answered “1” or “2” were categorized as “improved” on both transition questions. Analogously, those who scored “3”, “4” or “5” were categorized as “not improved” on the transition questions. We subsequently assessed the sensitivity to change for each core set item using two different procedures depending on whether the core set item was a binary vs. a continuous variable. For binary variables, we tested whether there was a statistically different distribution (at predefined  $p \leq 0.20$ ) among the patients who changed status from baseline to follow-up (e.g. no organ involvement at baseline to new involvement at follow-up, and vice versa) between the improved and non-improved patients with respect to either transition question using Fisher’s exact test. Core set items with significantly different distribution between the improved and non-improved groups with respect to either transition question were classified as sensitive to change.

For continuous variables, we assessed sensitivity to change using the effect size in the improved and non-improved groups. The effect size was defined as the change in the core set item from baseline to follow-up in each corresponding subgroup (e.g. in the improved and in the non-improved group) divided by the standard deviation at baseline for that subgroup. Continuous core set items for which the effect size was at least 0.20 (defined as small effect size) in the improved group with respect to either transition question was deemed as sensitive to change.

For both continuous and discrete core set items, sensitivity to change was determined using only the available data. Missing data was not imputed.

After all these attributes for each core set item were obtained, eight members of the steering committee underwent a modified Nominal Group Technique to determine which of the 31 core set items should be retained for the next step in the algorithm. Ed Giannini moderated this internet-based session using webinar. Specifically, steering committee members were

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

provided with the above-mentioned information (e.g. correlation values among the core set items, sensitivity to change results for each core set item, etc.) along with summary statistics for each of the 31 core set items. The Steering Committee members reviewed the data and scored each core set item on an ordinal scale (1-4) for the following attributes: feasibility, reliability, and face, content, and construct validity [including sensitivity to change] using the modified content validity index matrix. A score of 4 (highest score) was assigned when the item referred to a value or an attribute well-established in the literature or through systematically obtained information; a score of 3 indicated a value or an attribute somewhat known and accepted, but that may need minor alteration or modification; a score of 2 indicated that the rater was unable to assess the attribute without additional information or research; and a score of 1 (lowest score) meant that the attribute should definitely not be used as a core set item. Experts could also assign “not applicable” if they were unfamiliar with an item or different aspects of feasibility, reliability, and validity for the item. Based on each attribute, the experts were asked if the core set item should be considered in the combined response index; score of 3 or 4 with  $\geq 75\%$  consensus was considered supportive of an individual core set item. Other core set items were removed from next step.

As patients were not involved in any step of the CRISS development, pain and vitality score from the SF-36 (to assess fatigue) was added to list of core set items in representation of the patient’s pain and fatigue.

At the end of this consensus exercise, 16 core set items were retained for the next step in the algorithm development.

**(iii) Development and ratings of representative patient profiles.**

We developed representative patient profiles by sampling patients from the CRISS cohort and from the Canadian Scleroderma Research Group (CSRG) database. Since the CRISS cohort included only patients with dcSSc, in selecting patients from the CSRG database, we limited ourselves to patients who had dcSSc, disease duration of less than 5 years (from 1<sup>st</sup> non-Raynaud’s sign or symptom) and complete data at baseline and follow-up on all the 15 core set items, excluding “patient skin interference” which was not measured in the CSRG cohort. This identified N=94 patients in the CSRG database.

To assess whether the CSRG database and the CRISS cohort could be considered as two samples from the same population, we tested whether there were significant differences among the CRISS cohort and the subset of patients in the CSRG database that satisfied the CRISS cohort enrollment criteria. We compared the 15 selected core set items in the two populations at baseline and

follow-up and found no significant differences between the two groups of patients.

To impute patient skin interference for the CSRG patients, we used the CRISS cohort and fit a regression model, separately, for baseline and follow-up data, with patient skin interference as the outcome variable and the remaining 15 core set items as the predictors. Using the estimated regression coefficients obtained from the two linear regression models, we predicted patient skin interference for all the CSRG patients at baseline and follow-up, respectively. At baseline, being male and HAQ-DI at baseline were the only significant ( $p < 0.05$ ) predictors of skin interference. At follow-up, the modified Rodnan Skin Score (MRSS), Breathing visual analog score (VAS), GI VAS, Digital ulcers VAS, and HAD-DI were significant predictors of skin interference.

Patient profiles developed using the CRISS cohort were obtained by selecting those patients in the CRISS cohort that had no missing data at baseline and follow-up (N=43). To these 43 patients, we added an additional 13 subjects for a total of 56 patient profiles developed using the CRISS cohort. Of these additional 13 subjects in the CRISS cohort, 5 had no missing data at baseline but had missing data for one core set item at the 12-months follow-up, 3 subjects had no missing data at baseline but missing data for two core set items at the 12-months follow-up, and 5 subjects had no missing data at follow-up but missing data on two core set items at baseline. We chose to use those 5 subjects who had no missing data at follow-up but missing data at baseline on two core set items rather than subjects who had complete data at baseline but only missing one core set item at follow-up (there were 14 subjects with these characteristics in the CRISS cohort) based on the type of core set items for which the data was missing (e.g. discrete vs. continuous, more easily predictable based on the other available data or not). To impute the missing data for the 13 CRISS cohort patients, we used two different strategies depending on the nature of the core set item.

If the core set item for which we had missing data at a given time point (baseline or follow-up) was binary, we imputed the binary missing value by fitting a logistic regression to the binary variable of interest with the remaining core set items at the given time point as predictors. We then predicted the missing values for the subjects with missing data either at baseline or follow-up using the formula for the probability of success in logistic regressions. In cases where the prediction was ambiguous, e.g. the predicted probability of a success for the missing binary core set item was close to 50%, we set the missing value equal to 0 or 1 based on what was the most prevalent outcome in a subset of subjects that (i) had similar demographic information as the patient(s) in consideration, and (ii) had similar values, in the core set items that were found to be significant predictors in the logistic regression, to those of the patient(s) in consideration.

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

For continuous core set items, we used a similar strategy replacing the logistic regression with a linear regression model. In other words, for each continuous core set item with missing data either at baseline or follow-up in the 13 additional CRISS cohort subjects, we fit a linear regression model separately for baseline and follow-up data where the outcome measure was the continuous core set item with missing data and the covariates were the remaining core set items. Missing values were imputed according to the linear regression model.

The 56 patient profiles created using the CRISS cohort and the 94 patient profiles developed using the CSRG database were randomly assigned for ratings to a panel of 54 scleroderma experts, located in North America (N=29), Europe (N=21) and Australia (N=4), and with different years of experience on SSc management, over 10 years (N=38) versus at most 10 years (N=16). Each expert was asked to rate 20 patient profiles and to answer the following two questions:

- 1. Do you think the patient has improved, stabilized, or worsened (or unable to tell) over 1-year?;*
- 2. How would you rank the three most important core set items that influenced your decision regarding change or stability?*

Forty experts rated the profiles and 14 did not provide any rating.

We examined the raters' answers in several steps. In the first stage, we examined responses to the first question and we claimed that consensus was reached when at least 75% of the experts provided the same answer for a patient profile (N=107). Patient profiles on which consensus was not reached but where there was a 60-74% agreement among the experts (N=43), were rated again by the Steering Committee, and discussed in detail using the Nominal Group Technique. After discussion, the Steering Committee casted its votes on these patient profiles (N=43) and the initial ratings were tabulated to derive the percentage of agreement among raters. If there was at least 75% agreement, we stated that consensus was reached. Eleven additional patient profiles out of the 43 reevaluated by the Steering Committee, got added to the initial 107 patients where consensus was reached, yielding a total of 118 patient profiles.

Using data relative to the 118 patient profiles, we examined the experts' reply to the following question:

*How would you rank the three most important core set items that influenced your decision regarding change or stability?*

We tabulated how many times each of the core set item was deemed to be the most important, the second most important and the third most important core set item in helping a scleroderma expert determining whether a patient was

improved, worsened or stabilized. We created a 16 x 3 frequency table that reported for each core set item these counts, resulting in sixteen 3-dimensional vectors. To determine whether we could identify groups of core set items that had similar counts for the number of times they were deemed to be the most useful, the second most useful and the third most useful item in influencing a rater decision, we performed a cluster analysis on the sixteen 3-dimensional vectors. Specifically, we assessed whether clusters of core set items existed by running a well-established clustering algorithm, the K-means algorithm, on the 16 x 3 frequency table. Since the K-means algorithm requests that the number of groups in which to cluster the 16 core set items be specified a priori, we determined the number of clusters by running the algorithm with  $K=1,2,\dots,15$  clusters. We selected the appropriate number of clusters in which to group the 16 core set items by looking at the degree of similarity within each cluster, which can be quantified using the within-clusters sum of squares. A lower within-clusters sum of squares is preferred as it indicates that the clusters are rather homogeneous within themselves but they are different from one another. The appropriate number  $K$  of clusters was chosen as the smallest number  $K$  such that there was not an appreciable difference in the level of homogeneity within clusters if the core set items were grouped in to  $K$  clusters vs  $K+1$ . The within-clusters sum of squares criteria indicated that the 16 core set item could be appropriately grouped into 2 clusters, each including core set item with a similar degree of helpfulness in rating patient profiles. The first cluster included MRSS, FVC% predicted, patient global assessment, physician global assessment, and HAQ-DI and the second cluster included all the remaining core set items.

**(iv) Development of response definition.**

We developed the response definitions using data on the 118 patient profiles where consensus was reached. We used the raters' consensus decision as the binary outcome variable, with 1 indicating that a patient had been rated as improved at the 1-year follow-up ( $N=56$ ) and 0 indicating that a patient had not been rated improved (these included  $N=29$  that were considered to be stable and  $N=33$  who were deemed worsened). We used the change in the 16 core set items from baseline to follow-up as the predictors in a suite of logistic regression models, starting from simple logistic regression models with only one predictor to a model including all 16 core set items.

The different logistic regression models were compared based on the area under the curve (AUC), sensitivity and specificity where the gold standard was the consensus-based improved/not improved rating. Additionally, we compared the various logistic regression models based on their respective predictive power, which we quantified using the Brier score, a generalization of the Mean Squared Error metric for the binary core set items. Specifically, if  $y_i$  represents the raters' consensus opinion on patient  $i$  with  $y_i = 1$  if the patient has been rated

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

as improved and  $y_i = 0$  if the patient has been rated as not improved, and  $p_i$  is the predicted probability that the patient is improved, obtained from the logistic regression model, the Brier score is defined as:

$$\text{Brier Score} = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2$$

Among all the logistic regression models considered – from models with each core set item individually as a predictor, to models with two core set items, etc. – the model that included change from baseline to follow-up in MRSS, FVC% predicted, physician global assessment, patient global assessment, and HAQ-DI was considered by the steering committee as the model with the best face validity. This model had a sensitivity of 0.9821 (95% CI: [0.9816, 0.9827]), specificity of 0.9310 (95% CI: [0.9300, 0.9321]), an AUC of 0.9861 and a Brier score was 0.038 (lower score indicates better predictive performance).

Having decided the core set items that are part of the response definition, we derived the predicted probability of improvement at 1-year follow up for each patient profile using the logistic regression model with the 5 core set items identified above. We tested whether there was a significant difference in the distribution of the probability of improvement at 1-year follow-up among the improved and not-improved subjects using the Kolmogorov-Smirnov test. The continuous predicted probability of improvement was translated into a binary classification of patients as “improved” and “not improved” using a cutoff  $c$  for the predicted probability. In other words, a patient will be deemed “improved” at the 1-year follow-up if the predicted probability of improvement for the patient is greater than  $c$ , otherwise he/she is considered “not-improved”. To establish the cutoff  $c$  to use to transform the predicted probabilities into an “improved”-“not improved” classification, we derived the specificity and sensitivity of the corresponding binary classification for various values of “ $c$ ”. By plotting the sensitivity and specificity of the classification as a function of the probability threshold “ $c$ ”, we determined the predicted probability cutoff that yielded the best balance of sensitivity and specificity as 0.6.



**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

**Supplementary Table 1. Correlation between the continuous core items among the 14 core items at baseline.\***

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
V1	1.0	-0.26	0.43	0.60	0.33	0.49	0.31	0.04	0.16	0.09	0.09	0.04	0.03	0.17
V2		1.0	-0.22	-0.33	-0.23	-0.20	-0.18	0.02	-0.03	-0.17	-0.003	-0.11	-0.27	-0.16
V3			1.0	0.46	0.57	0.66	0.56	0.23	0.26	0.17	0.02	-0.06	0.28	0.25
V4				1.0	0.45	0.54	0.33	0.17	0.18	0.11	0.04	0.08	0.13	0.10
V5					1.0	0.55	0.57	0.35	0.35	0.19	-0.02	0.01	0.41	0.30
V6						1.0	0.60	0.19	0.44	0.26	0.11	0.06	0.30	0.22
V7							1.0	0.17	0.47	0.41	0.11	0.09	0.34	0.33
V8								1.0	0.15	0.06	-0.05	0.06	0.26	0.07
V9									1.0	0.35	0.20	0.15	0.39	0.45
V10										1.0	0.16	0.11	0.20	0.23
V11											1.0	-0.04	-0.02	0.02
V12												1.0	0.19	0.07
V13													1.0	0.36
V14														1.0

V1=MRSS, V2=FVC% predicted, V3=HAQ-DI, V4=Physician global, V5=Patient global, V6=Patient skin interference, V7=Pain, V8=Vitality, V9=Raynaud VAS, V10=Digital Ulcers VAS, V11=Number of digital ulcers, V12=BMI, V13=Breathing VAS, V14=GI VAS

\*renal crisis and tendon friction rubs not included

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

**Supplementary Table 2. Correlation between the change scores in the 14 core continuous core items.\***

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
V1	1.0	-0.30	0.22	0.26	0.16	0.32	0.21	0.12	0.17	0.17	-0.10	0.07	0.08	0.17
V2		1.0	-0.39	-0.31	-0.27	-0.29	-0.33	0.03	-0.06	-0.17	0.10	0.002	-0.30	-0.10
V3			1.0	0.17	0.27	0.31	0.23	-0.005	0.08	-0.05	-0.009	-0.18	0.30	0.05
V4				1.0	0.25	0.46	0.19	-0.09	0.18	0.03	-0.08	0.04	0.33	0.26
V5					1.0	0.13	0.25	-0.007	0.002	0.05	-0.14	-0.10	0.16	0.25
V6						1.0	0.28	-0.08	0.15	-0.07	-0.02	0.22	0.30	0.02
V7							1.0	0.07	0.27	0.10	0.22	0.11	0.33	0.23
V8								1.0	0.001	-0.12	-0.03	0.01	-0.12	-0.14
V9									1.0	0.20	0.35	0.20	0.23	0.47
V10										1.0	-0.13	0.11	0.05	0.36
V11											1.0	0.008	0.06	0.05
V12												1.0	0.16	-0.07
V13													1.0	0.28
V14														1.0

V1=MRSS, V2=FVC% predicted, V3=HAQ-DI, V4=Physician global, V5=Patient global, V6=Patient skin interference, V7=Pain, V8=Vitality, V9=Raynaud VAS, V10=Finger Ulcers VAS, V11=Number of digital ulcers, V12=BMI, V13=Breathing VAS, V14=GI VAS

\*renal crisis and tendon friction rubs not included

ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology

**Supplementary Table 3. Example of a patient rated “improved” by the experts. Predicted probability of improving is 0.99 according to CRISS.**

	Baseline	Follow-up	Absolute change
Age	51.6 years		
Disease duration (months)	12.98		
<b>Global assessments</b>			
<i>Patient global assessment (0-10)*</i>	3	1	-2
<i>Physician global assessment (0-10)*</i>	3	3	0
<b>Musculoskeletal</b>			
<i>HAQ-DI (0-3)*</i>	0.625	0	-0.625
Tendon friction rubs*	No	No	No change
<b>Skin</b>			
<i>MRSS (0-51)*</i>	13	3	-10
Patient skin interference last month	2	0	-2
<b>Lung</b>			
<i>FVC% predicted*</i>	62	75	13
Breathing VAS (0-10)	2	0	-2
<b>Renal</b>			
<i>Renal crisis**</i>	No	No	No change
<b>Gastrointestinal</b>			
GI VAS (0-10)	3	3	0
Body Mass Index (BMI)	25.40	26.58	1.18
<b>Raynaud's</b>			
Raynaud's VAS (0-10)	2	1	-1
<b>Digital ulcers</b>			
Digital ulcers VAS (0-10)	0	0	0
Number of digital ulcers	0	0	0
<b>HRQOL</b>			
Pain VAS (0-10)	3	1	-2
Fatigue (SF-36 Vitality scale) (0-100)	42.31	35.12	-7.19

\*included in Step 2; \*\* included in Step 1

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

**Supplementary Table 4. Example of a patient rated “improved” by the experts. Predicted probability of improving is 0.60 according to CRISS.**

	Baseline	Follow-up	Absolute change
Age	64.65 years		
Disease duration (months)	30.74		
<b>Global assessments</b>			
<i>Patient global assessment (0-10)*</i>	1	0	-1
<i>Physician global assessment (0-10)*</i>	7	4	-3
<b>Musculoskeletal</b>			
<i>HAQ-DI (0-3)*</i>	0.375	0.250	-0.125
Tendon friction rubs*	No	No	No change
<b>Skin</b>			
<i>MRSS (0-51)*</i>	21	15	-6
Patient skin interference last month	8	5	-3
<b>Lung</b>			
<i>FVC% predicted*</i>	86	81	-5
Breathing VAS (0-10)	0	0	0
<b>Renal</b>			
<i>Renal crisis**</i>	Yes	Yes	No change
<b>Gastrointestinal</b>			
GI VAS (0-10)	0	0	0
Body Mass Index (BMI)	25.12	24.82	-0.3
<b>Raynaud's</b>			
Raynaud's VAS (0-10)	3	4	1
<b>Digital ulcers</b>			
Digital ulcers VAS (0-10)	0	8	8
Number of digital ulcers	0	0	0
<b>HRQOL</b>			
Pain VAS (0-10)	0	2	2
Fatigue (SF-36 Vitality scale) (0-100)	35.12	35.12	0.0

\*included in Step 2; \*\* included in Step 1

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

**Supplementary Table 5. Example of a patient rated “worsened” by the experts. Predicted probability of improving is 0.002 according to the CRISS.**

	Baseline	Follow-up	Absolute Change
Age	53.6 years		
Disease duration (months)	43.3		
<b>Global assessments</b>			
<i>Patient global assessment (0-10)*</i>	1	2	1
<i>Physician global assessment (0-10)*</i>	1	2	1
<b>Musculoskeletal</b>			
<i>HAQ-DI (0-3)*</i>	0	0	0
Tendon friction rubs*	No	Yes	Change to worsen
<b>Skin</b>			
<i>MRSS (0-51)*</i>	7	5	-2
Patient skin interference last month	3	2	-1
<b>Lung</b>			
<i>FVC% predicted*</i>	87	80	-7
Breathing VAS (0-10)	0	1	1
<b>Renal</b>			
<i>Renal crisis**</i>	No	No	No change
<b>Gastrointestinal</b>			
GI VAS (0-10)	0	1	1
Body Mass Index (BMI)	24.68	24.68	0
<b>Raynaud's</b>			
Raynaud's VAS (0-10)	0	3	3
<b>Digital ulcers</b>			
Digital ulcers VAS (0-10)	0	0	0
Number of digital ulcers	0	0	0
<b>HRQOL</b>			
Pain VAS (0-10)	1	1	0
Fatigue (SF-36 Vitality scale) (0-100)	37.52	35.10	-2.42

\*included in Step 2; \*\* included in Step 1

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

**Supplementary Table 6. One core item logistic model using expert consensus definition of improved vs. not**

Core item	Area under the curve (AUC)	Sensitivity	Specificity	Brier Score
MRSS	0.9231	0.8392	0.8793	0.108
FVC% predicted	0.7906	0.6429	0.7586	0.184
Physician global	0.7743	0.7143	0.7241	0.197
Patient global	0.7448	0.7143	0.6207	0.204
HAQ-DI	0.7107	0.6429	0.6897	0.200
Pain	0.6857	0.6071	0.7586	0.218
Vitality	0.6856	0.4643	0.7414	0.225
VAS Breathing	0.6670	0.375	0.8103	0.219
GI VAS	0.6667	0.7857	0.4483	0.220
Patient skin interference last month	0.6601	0.5179	0.7586	0.226
Raynaud's VAS	0.6190	0.4286	0.7241	0.238
Tendon friction rubs	0.5640	0.2321	0.8966	0.245
Digital ulcers VAS	0.5503	0.2857	0.7931	0.247
Body mass index	0.4946	0.1786	0.8276	0.250
Number of digital ulcers	0.4764	0.0179	0.931	0.249

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity, GI= gastrointestinal, VAS= visual analog scale

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

**Supplementary Table 7. Two core item logistic model using expert consensus definition of improved vs. not**

Core item	Area under the curve (AUC)	Sensitivity	Specificity	Brier Score
MRSS, FVC% predicted	0.9632	0.8929	0.9138	0.068
MRSS, HAQ-DI	0.9615	0.9107	0.8793	0.076
MRSS, Patient global	0.9560	0.875	0.8966	0.081
MRSS, physician global	0.9450	0.875	0.9310	0.094
FVC% predicted, HAQ-DI	0.8519	0.7679	0.8448	0.158
FVC% predicted, Patient global	0.8548	0.7679	0.8448	0.152
FVC% predicted, physician global	0.8544	0.750	0.8103	0.158
HAQ-DI, patient global	0.7982	0.7143	0.7241	0.184
HAQ-DI, physician global	0.8094	0.6607	0.7931	0.181
Patient global, physician global	0.8265	0.7321	0.7759	0.170

HAQ-DI= health assessment questionnaire-disability index, MRSS= modified Rodnan skin score, FVC= Forced vital capacity

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

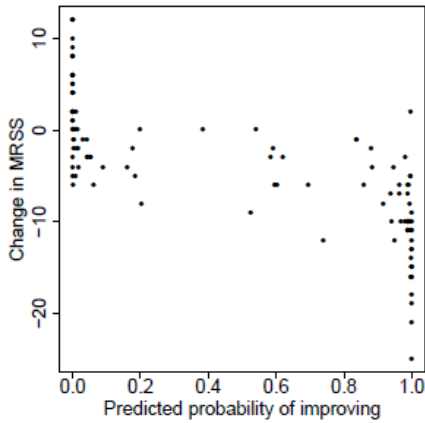


**Supplementary Table 8. Definition of scleroderma renal crisis (adapted, with permission, from Steen VD, Mayes MD, Merkel PA. Assessment of kidney involvement. Clin Exp Rheumatol 2003;21 Suppl 29:29–31).**

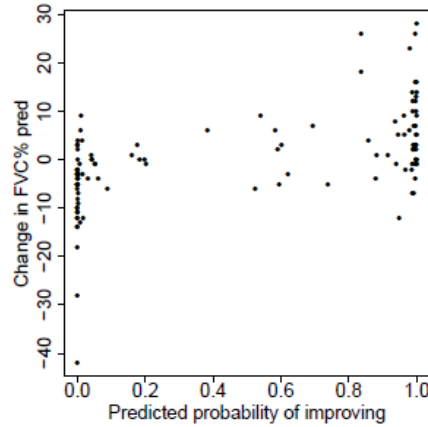
<i>A. Hypertensive SRC (fulfills both A1 and A2)</i>
1. New onset hypertension, defined as any of the following:
a) Systolic blood pressure $\geq 140$ mmHg
b) Diastolic blood pressure $\geq 90$ mmHg
c) Rise in systolic blood pressure $\geq 30$ mmHg
d) Rise in diastolic blood pressure $\geq 20$ mmHg
AND
2. One (1) of the following five (5) features:
a) Increase in serum creatinine by 50+% over baseline OR serum creatinine $\geq 120\%$ of upper limit of normal for local laboratory
b) Proteinuria $\geq 2+$ by dipstick
c) Hematuria $\geq 2+$ by dipstick or $\geq 10$ RBCs/HPF
d) Thrombocytopenia: $< 100,000$ platelets/ $\text{mm}^3$
e) Hemolysis defined as anemia not due to other causes and either of the following: (1) Schistocytes or other RBC fragments seen on blood smear (2) increased reticulocyte count
<i>B. Normotensive SRC (fulfills both B1 and B2)</i>
1. Increase in serum creatinine $> 50\%$ over baseline OR serum creatinine $\geq 120\%$ of upper limit of normal for local laboratory
AND
2. One (1) of the following five (5) features:
a) Proteinuria $\geq 2+$ by dipstick
b) Hematuria $\geq 2+$ by dipstick or $\geq 10$ RBCs/HPF
c) Thrombocytopenia: $< 100,000/\text{mm}^3$
d) Hemolysis defined as anemia not due to other causes and either of the following: (1) Schistocytes or other RBC fragments seen on blood smear (2) Increased reticulocyte count
e) Renal biopsy findings consistent with scleroderma renal crisis (microangiopathy)

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

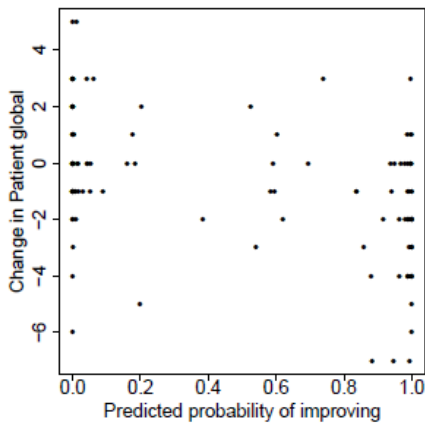
Supplementary Figure 1. (a) Change in MRSS, (b) Change in FVC% predicted, (c) Change in patient global assessment, (d) Change in physician global assessment, and (e) Change in HAQ-DI versus the predicted probability of improving yielded by CRISS.



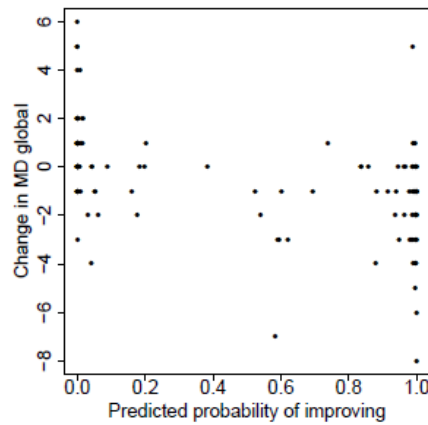
(a)



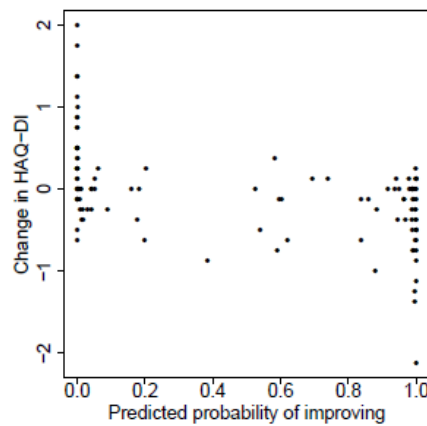
(b)



(c)



(d)



(e)

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

**ACR provisional composite response index for clinical trials in early diffuse cutaneous systemic sclerosis. Khanna D et al Arthritis and Rheumatology**

Supplementary Figure 2. Histogram of the predicted probabilities of improving in subjects in the RCT study of methotrexate vs. placebo.

